KNN 분류기의 범주할당 방법 비교 실험

A Comparative Study on Category Assignment Methods of a KNN Classifier

이영숙, 정영미 (연세대학교 문헌정보학과)

Young-Sook Lee, Young-Mee Chung
Dept. of Library and Information Science, Yonsei University

KNN(K-Nearest Neighbors)을 사용한 문서의 자동분류에서는 새로운 입력문서에 범주를 할당하기 위해 K개의 유사문서로부터 범주별 문서의 분류빈도나 유사도를 이용한다. 본 연구에서는 KNN 기법에서 보편적으로 사용되는 범주 할당 방법을 응용하여 K개 유사문서 중 최상위 및 상위 M개 문서에 가중치를 부여하는 방법들을 고안하였고 K값의 변화에 따른 이들의 성능을 비교해 보았다.

1 서론

전자문서의 증가는 필요한 정보를 신속하고 다양하게 접할 수 있도록 하는데 기여해 왔다. 그러나 이용자들의 수요가 증가하고 요구가 복잡, 다양해지면서 여러 검색엔진이나 정보검색 관련 기관에서는 좀더 정확하고 필요한 정보만을 검색할 수 있는 세련되고 고급화된 방법을 개발하기 위해 고심하고 있다.

이러한 상황에서 해결책의 하나로 제시되고 있는 것이 범주정보의 이용이다. 웹 로봇이 검색해 온 다양한 전자문서들에 적절한 범주를 할당하여 범주별로 저장해 둔다면, 단순히 단어매칭을 이용하는 것보다 적합한 문서를 찾을 가능성이 더 높아질 수 있게 된다.

자동 문서 범주화는 먼저 범주가 미리 할당된 문서들로 구성된 학습문서가 주어지고 자질 추출과정을 통해 각 문서에서 자질이 출현한 빈도를 계산한다. 자질-문서 출현빈도를 구한 후에는 특정 학습기법을 이용해 범주를 학습하고 범주를 할당할 새로운 문서가 입력되면 학습문서에 적용한 과정과 동일한 자질과 가중치의 집합으로 표현한 후, 범주 결정과정을 거쳐 적합한 범주를 할당한다.

기계학습을 이용해 문서에 범주를 자동으로 할당하는 자동 범주화 기법은 학습 및 범주결정방법에 따라 크게 규칙기반방법, 예제기반방법, 통계기반방법, 신경망 기반 방법 등으로 나누어진다.

K-최근접 문서(K-Nearest Neighbors:KNN) 기법은 대표적인 예제기반 범주화 기법이다. 학습문서들로부터 자질을 추출하고 학습문서 및 각 학습문서에 부여된 범주들을 벡터로 표현한 후, 입력문서와 가장 유사한 K개의 학습문서를 찾아 그 문서들에 이미 할당된 범주정보를 이용하여 입력문서의 범주를 결정한다.

KNN분류기는 여러 문서 범주화 기법 중에서도 알고리즘이 비교적 단순하여 다양한 분야에서 응용되고 있다. 그러나 새로운 입력문서와의 유사도가 높은 K개 학습문서들을 이용한 범주 할당시 '최적의 범주'를 할당할 수 있는 방법에 관한 연구는 부족하다. 따라서 본 연구에서는 KNN분류기의 특성을 고려한 8가지 알고리즘을 사용하여 KNN 분류기의 범주 할당 성능을 비교하고자 한다.

2 KNN의 범주 할당 방법 비교 실험

2.1 실험 배경

KNN을 이용한 문서의 자동분류는 주로 입력문서와 가장 유사한 K개 문서들의 유사도 또는 범주빈도(각 범주당 학습문서의 분류빈도, 한 학습문서가 해당범주에 분류된 경우 1, 아니면 0)를 합산(총 분류빈도)하여 그 값이 높은 범주를 시스템이 지정한 수 만큼 차례대로 입력문서의 범주로 할당하는 방법을 사용하고 있다. 따라서 얼마나 유사한 문서들을 찾아내며 그 문서들의 정보를 유용하게 이용하느냐에 따라 최적 범주 할당여부가 결정된다.

본 연구에서는 이러한 KNN 분류기의 특성을 이용하여 입력문서와 유사한 학습문서들의 정보를 더욱 활용할 수 있는 방향으로 4가지 알고리즘을 고안(m5,m6,s1,s2)하여 기존의 실험에서 사용된 방법들(m1,m2,m3,m4)과 비교해 보았다.

2.2 실험방법

본 실험에서 이용된 실험집단은 KTSET이며, 한글문서 254건의 제목과 초록을 대상으로 학습문서 174건, 실험문서 80건으로 구분하였다. 그리고 KTSET에서 사용하고 있는 ACM 분류체계중 B-H에 속하며 범주당 해당문서수가 2-60인 범주 180개를 학습에 이용하였다.

자질선정을 위해서는 문서빈도(Document Frequency: DF)를 이용하였다. DF 3 이상의 문서만을 대상으로 자질을 선정한 결과 자질수가 약 79%감소되어 총 751개의 자질이 추출되었다.

선정된 자질들을 벡터로 표현하기 위해 'TF(단어빈도) * IDF(역문서빈도)'를 가중치로 사용하였으며 학습문서(D_j)와 입력문서(D_x)의 유사도를 계산하기 위해 다음과 같은 코싸인 유사계수공식을 이용하였다.

$$sim(D_x, D_j) = \frac{\sum t_{xk} \times t_{jk}}{\sqrt{\sum (t_{xk})^2} \times \sqrt{\sum (t_{jk})^2}}$$

t_{xk}, t_{jk} : D_x, D_j 벡터에 출현한 용어K의 가중치

코싸인 유사계수를 이용하여 입력문서와의 유사도순으로 K개의 학습문서를 추출한다. 이때, K개 학습문서 각각에 부여된 범주들은 입력문서에 할당될 후보 범주리스트가 된다. 이들 후보 범주리스트로부터 입력문서에 할당할 최적의 범주를 찾아 내기 위해, 각 범주에 학습문서가 분류된 빈도수의 총합이나 유사도를 이용한 범주별 적합성 점수를 구한다. 그리고 높은 점수를 받은 상위 범주들을 입력문서에 할당한다.

2.3 범주 할당 방법

KNN을 이용한 범주 할당시, 입력문서와 유사한 학습문서들의 유사도와 범주빈도를 다양하게 응용하여 보다 우수한 성능의 범주할당방법을 찾아내기 위해 다음과 같은 8가지 방법들의 성능을 비교하였다.

기존의 KNN실험에서 주로 사용하는 범주할당 방법	K개의 문서중 최상위 문서에만 가중치를 주는 방법	K개의 문서중 상위 M개의 문서에 가중치를 주는 방법
m1,m2, m3, m4	m5, m6	s1, s2

[1] m1: ND(범주를 할당할 새로운 입력문서)와 유사도가 높은 K개 문서중 최상위 학습문서에 속하는 범주를 ND의 범주로 모두 할당

[2] m2: K개 문서들의 범주빈도를 이용하며, 각 범주별로 학습문서들이 분류된 빈도를 합산(총 분류빈도)하여 그 값이 높은 범주를 ND에 할당

[3] m3: K개 문서들 각각의 유사도를 학습문서(D_j)가 해당범주(C_k)에 속할 조건확률($P(C_k|D_j)$)과 곱하고 이들을 모두 합산하여 각 범주별로 적합성 점수($rel(C_k|D_x)$)를 구함. 그런 다음 적합성점수가 높은 상위 범주를 ND에 할당(Yang 1994, 16)

$$rel(C_k|D_x) = \sum_{D_j \in \{K개의 최상위 문서\}} sim(D_x, D_j) \times P(C_k|D_j)$$

$$P(C_k|D_j) = \frac{\text{범주 } C_k \text{가 문서 } D_j \text{에 할당된 빈도}}{\text{학습집단에서 문서 } D_j \text{가 출현하는 빈도}}$$

[4] m4: K개 문서들의 범주 빈도 합산으로 범주리스트를 만든 후, 동점이 있는 경우 m3와 같은 적합성 점수를 이용(m2 + m3)

[5] m5: K개 문서들 중 최상위 학습문서 한 건(K_1)의 범주빈도에 가중치(W_{frq})를 더한 후, m2와 같은 방법으로 K개 문서들의 범주빈도를 이용하여 각 범주별로 학습문서들의 총 분류빈

도를 구하고 그 값이 높은 상위 범주를 ND에 할당

*W_{frq}=1

[6] m6: K개 문서들 중 최상위 학습문서 한 건(K_1)의 유사도에 가중치(W_{sim})를 더한 후, m3와 같은 방식으로 적합성 점수를 범주별로 구하고 그 값이 높은 상위 범주를 ND에 할당

*W_{sim}= (Sim_{K1} / K)+1

(Sim_{K1}: 입력문서와 가장 유사도가 높은 학습문서(K_1)의 유사도 값)

[7] s1: K개 문서들 중 상위 M개 문서의 범주빈도에 순위정보를 이용한 가중치(W_{frq})를 더한 후, m2와 같은 방법으로 범주별 문서의 총 분류빈도를 구하여 그 값이 높은 상위 범주를 ND에 할당 (범주빈도 합산후 동점이 생기는 경우는 m4와 같이 유사도순으로 순위를 부여)

*W_{frq}= 1 / Rank

(Rank: 유사도순으로 정렬된 K개 문서들의 순위(1,2,3,...,M,...K))

[8] s2: K개 문서들 중 상위 M개 문서의 유사도에 순위정보를 이용한 가중치(W_{sim})를 더한 후, m3와 같은 방식으로 적합성 점수를 구하여 그 값이 높은 상위 범주를 ND에 할당

$$*W_{sim} = \frac{(Sim/Rank)}{K}$$

3 실험결과

3.1 평가방법

8가지 방법들의 성능을 비교하기 위한 평가척도로 할당성공율, 재현율, 정확률을 이용하였다. 할당성공율은 실험문서에 할당된 범주가 KTSET에서 미리 부여된 분류체계와 일치하는지의 여부만 판단하여 실험문서 당 할당된 적합 범주가 있는 경우 그 수에 상관없이 무조건 1을 주고 할당된 범주가 없을 경우 0을 주었다. 재현율과 정확률은 정보검색연구에서 많이 사용되는 평가기법이며 공식은 다음과 같다 (Yang1998,6).

$$\text{재현율} = \frac{\text{시스템에의해할당된범주중적합범주의수}}{\text{적합범주의총수}}$$

$$\text{정확률} = \frac{\text{시스템에의해할당된범주중적합범주의수}}{\text{시스템에의해할당된범주의총수}}$$

3.2 결과

K값은 1부터 100까지 변화시키고, s1과 s2에 가중치를 부여하기 위한 M은 K가 5이상일 때부터 M=K/2 만큼의 수를 적용하였다. K가 전체 학습문서의 약 30%지점(K=50)에서부터는 s2를 제외한 모든 방법들의 할당성공율, 재현율, 정확률 성능이 낮아졌다.

<그림 1>을 보면 K가 전체 학습문서의 16-17%(K=28-30)일 때 s1이 최고할당성공율을 보여주고 있으며 특히, 20%이하에서 8가지 방법 중 가장 좋은 성능을 보여주고 있다. m1의 경우, 이미 Yang(1994, 17)의 연구에서도 낮은 성능이 입증되었다. 유사도를 이용한 m3, m6, s2는 모두 우수한 성능을 보여주고 있으며, 특히 상위문서들에 가중치를 부여한 s2는 가장 안정된 성능을 보여주었다. 범주빈도를 이용한 m2, m4, m5는 유사도를 이용한 방법들에 비해 낮은 성능을 보여주고 있으나 최상위 문서에 가중치를 부여한 m5의 경우는 그 중 나은 성능을 나타내었다.

<그림 2>와 <그림 3>의 재현율과 정확률 곡선을 보면, 두 경우 모두 K=18일 때 s1, K=23일 때 m3에서 각각 가장 우수한 성능을 보여준다. 그리고 K가 전체 학습문서의 8-10 % 이하일 때는 s1이 꾸준하게 좋은 성능을 보여주었고 s2는 m3와 비교했을 때 K값의 변화에 상관없이 안정되고 높은 성능을 보여주었다.

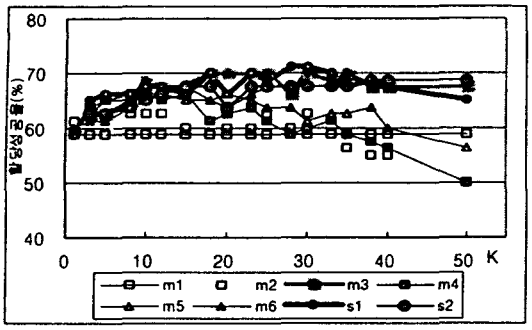
<표 1>은 K값에 따라 변화하는 각 범주할당방법들의 성능을 평균한 값이다. 할당성공율에서는 m3, 재현율과 정확률에서는 s2가 가장 높은 평균값을 보여주며 m3보다 좋은 성능을 보여주는 것을 알 수 있다. 또한 s1의 경우, K가 좋은 성능을 보이다가 10%이상에서부터 급격히 낮아지므로 전체적인 평균값이 m3보다 낮다.

선행연구들이 범주빈도의 합산(m2)보다 유사도를 이용한 범주할당(m3)을 더 선호하는 이유는 m2가 동점(ties)이 많이 나오기 때문이다 (Larkey and Croft 1996). 이러한 점을 보완하기 위하여 고안된 m4는 예상외로 낮은 성능을 보였다.

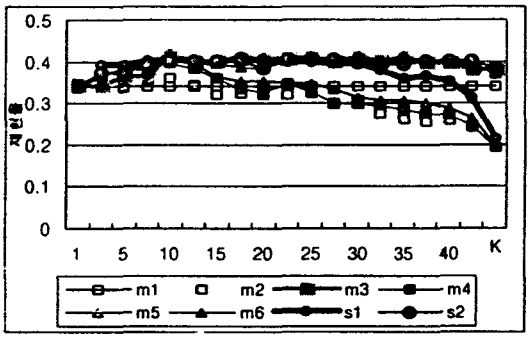
4 결론

KNN분류기를 이용한 선행연구들은 범주빈도를 이용하기도 하지만, 대부분, K개 학습문서들의 유사도를 이용하여 범주를 할당하였다.

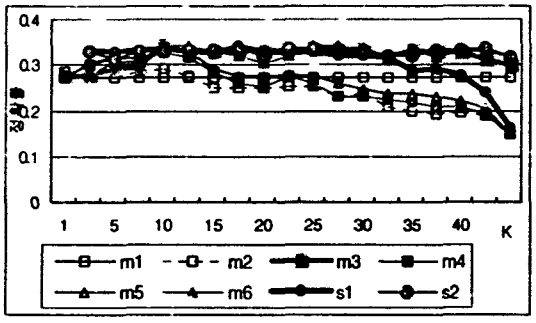
본 연구에서는 이를 응용하여 K개 학습문



<그림 1> K 값의 변화에 따른 할당성공율



<그림 2> K값의 변화에 따른 재현율



<그림 3>K값의 변화에 따른 정확률

<표 1>K값에 따른 각 범주할당방법들의 성능평균

평균	할당성공율(%)	재현율	정확률
m1	58.75	0.3385	0.2708
m2	58.68	0.3047	0.2386
m3	66.97	0.3914	0.3208
m4	60.26	0.3211	0.2541
m5	62.23	0.3335	0.2653
m6	65.92	0.391	0.3203
s1	66.80	0.3766	0.3074
s2	66.52	0.3947	0.3275

서 중 유사도순으로 최상위이거나 상위 M개에 속하는 문서에 가중치를 주는 알고리즘을 고안하였고 기존의 방법들과 비교해 보았다.

실험결과, 각 범주당 학습문서의 총 분류빈도보다는 유사도를 이용하는 것이 더 좋은 성능을 보여주었다. 그리고 K개 문서 전체를 이용하기 보다는 상위문서의 정보를 활용하는 것이 더욱 바람직함을 알 수 있었다. 특히 기존의 KNN응용연구에서 주로 사용된 m3방법과 본 실험에서 고안한 s1, s2를 비교했을 때, s1은 K값이 30 이하일 때 m3보다 더 우수한 성능을 보여주었고, s2는 K값에 상관없이 우수하고 안정된 성능을 보여주었다. 또한 최상위 문서의 유사도에 가중치를 더해준 m6또한 비교적 안정된 성능을 보여주었다. 이는 K개의 유사문서 중 비교적 높은 유사도값을 가지는 상위 학습문서를 가중치를 통해 적극적으로 활용함에 따라 입력문서의 범주할당에 도움을 주지 못하는 범주들, 즉 노이즈(noise)정보가 범주할당 과정에 미치는 영향을 최소화하기 때문인 것으로 분석된다. 결론적으로, K값이 소규모일 때는 s1을, K값이 클 경우에는 s2를 사용하는 것이 성능면에서 바람직하다고 할 수 있다.

그러나 자동 범주 할당을 위한 범주를 학습시키기에 범주당 문서수(평균 5)와 문서당 범주수(평균 2)가 부족하며, 이로 인해 전체적으로 성능이 높지 않은 것이 본 연구의 한계로 지적될 수 있다. 따라서 범주당 문서수, 문서당 범주수에 따른 성능의 차이 및 다양한 M값에 따른 s1과 s2의 성능측정에 관한 연구가 뒤따라야 하겠다.

참고문헌

Larkey, L.S. and W.B. Croft. 1996. "Combining Classifiers in Text Categorization". SIGIR' 96, 287-297.

Yang, Y. 1994. "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorizatin- and Retrieval". SIGIR'94, 13-22.

Yang, Y. 1998. "An Evaluation of Statistical Approaches to Text Categorization". INRT Journal, 1-19.