Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 30 - May 1, 2004, pp. 195-195

나이브 베이즈와 군집분석을 결합한 전자메일 필터

(An Email Filter combining Naive Bayes and Cluster Analysis)

신양규[1]

초 록

전자메일의 필터링 기법들 중 주로 많이 사용되는 나이브 베이즈는 메일에 나타난 단어들로 특성벡터를 구성할 때 단어의 출현 순서는 무관하다는 조건부 독립 가설을 적용한다. 이와 같은 특성 때문에 나이브 베이즈 기법을 사용하는 전자메일 필터의 정확율은 현재까지 97%라는 한계를 극복하기 힘들었다. K-평균 군집분석기법은 주어진 전자메일을 범주별로 나눌 수 있기 때문에, 각 범주 중심에서 적정 반경을 벗어나는 전자메일들을 대상으로 나이브 베이즈 정리를 적용시켜 필터링할 범주에의 소속 여부를 결정할 수 있다. 이 방법의 장점은 K-평균 군집분석을 적용할 때 모수 K를 자유롭게 조절할 수 있으므로, 임계값을 결정하기 위해 사용하는 사람의 주관적 판단을 적용할 수 있다는 점이다. 현재까지의 실험 결과는 나이브 베이즈와 군집분석기법을 결합한 방법에서 98% 이상의 정확율을 보임으로서 판별력의 개선이 있음을 보여주고 있다.

[1]대구한의대학교 자산운용과학과