Proceedings of the Korean Data And Information Science Society,
October 22-23, 2004, pp. 95-95

카이제곱 통계량을 이용한 메일 필터
(An Email filter using Chi-square statistics)

신양규[1)]

초록

스팸 메일과 같이 원하지 않는 메일을 탐지하여 삭제하거나 지정된 곳에 저장하는 메일 필터는 주로 나이브 베이즈 방법을 사용하고 있다. 그러나 나이브 베이즈 방법은 메일처럼 단어 위주의 문서를 분류하는데 정확율의 향상에 한계가 있다는 단점을 가지고 있다. 본 논문에서는 카이제곱 통계량을 이용하여 미리 분류된 메일 집합의 성질을 파악하고, 이를 바탕으로 새로운 메일이 주어졌을 때 파악된 성질을 만족하는지를 검증하도록 하였다. 카이제곱 통계량을 활용한 검증에서는 주어진 메일이 가지는 단어의 갯수에 따라 자유도를 설정하였다. 나이브 베이지 방법을 활용한 분류법과 카이제곱 통계량을 활용한 분류법의 정확율을 비교하기 위해 같은 메일 집합을 대상으로 나이브 베이즈 방법과 카이제곱 통계량 방법을 병행하여 분석하였다. 메일에서 단어를 기초 데이터로 추출하여 두 가지 분류법을 비교한 결과 전체 약 3,000개의 검증 데이터에서 나이브 베이즈 방법이 87.9%, 카이제곱 방법이 92.6%로 나타나 카이제곱 통계량을 활용한 방법이 우수한 것으로 볼 수 있다.

1 경상북도 경산시 유곡동 대구한의대학교 자신운용과학부 교수. yks@dhu.ac.kr