Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 73-73

문서 데이터 군집화에 대한 연구

신양규[1]

요 약

인터넷과 정보통신 기술의 발전으로 전자저널, 뉴스그룹, 전자우편, 웹 문서, 그리고 각종 업무용 문서 등 다양한 전자 문서들이 폭증하고 있는 상황에서 이들 문서를 유사한 것들끼리 군집화할 필요성이 점차 증가하고 있다. 하지만 문서들의 군집화는 다변량 분석에서 가정하는 일반적인 군집화 환경과 중요한 차이점이 있다. 즉, 문서를 하나의 개체로 볼 때 문서 개체가 가지는 속성인 단어의 수가 문서마다 일정하지 않다는 점이다. 따라서 문서의 분류를 위해 일반적으로 사용하는 벡터공간 모델은 주어진 문서 전체의 집합을 고차원의 희소행렬로 표현하는데, 이 행렬에는 문서에서 의미를 가지는 단어들만을 추출하여 속성으로 설정한 후 각 문서들을 해당 속성의 빈도수를 값으로 가지는 열벡터로 설정된다. 이와 같은 방법으로 행렬을 문서 검색에 적용한 연구로 Berry, Dumais & O'Brien(1995)은 singular value decomposition을 이용하여 단어와 문서 사이에 서로 연관된 잠재적인 관계를 찾는 방법을 제안하였고, Kolda(1997)는 행렬의 근사법(approximation)을 효율적으로 계산하는 방법을 제시하였으며, Papadimitriou, Raghavan, Tamaki & Vempala(1998)은 확률적 프로젝션을 이용한 행렬 근사법을 연구하였다. 또한 Dhilon & Modha(2000) 와 Dhillon, Fan & Guan(2001)은 행렬을 이용한 대규모 문서의 군집화 방법으로 구면형 k-평균 군집화 기법을 제안하였다. 문서의 군집화에 사용하는 기법들은 대부분 반복 계산을 통한 수렴기법의 응용으로 볼 수 있는데, 이들은 많은 수의 지역 극소점들(local minima) 중 어느 하나로 수렴시키게 된다. Bradley와 Fayyad(1988)는 주어진 초기 입력에서 더욱 정교한 시작 조건을 생성하는 방법을 제안하였는데, 이 방법은 분포의 형태를 추정하는 효율적인 기법에 바탕을 두고 있으며, 생성된 정교한 시작 조건은 반복 알고리즘을 통해 더 나은 지역 극소점으로 수렴하게 된다.

본 연구에서는 텍스트 문서들을 내용이 유사한 것들끼리 군집화하기 위해 각 군집의 개념벡터를 생성하였다. 개념벡터는 군집 내에 속한 각 텍스트 문서에 대응하는 벡터들의 중심 벡터로 계산하였으며, 벡터들은 단위벡터로 정규화 되어 n차원 구면 상에서 군집화가 이루어진다. 개선된 k-평균 군집분석의 효율성과 정확성을 입증하기 위해 Classic3라는 문서 데이터 집합을 대상으로 실험하였으며, 결과는 같은 정확도에서 7% 이상 처리 속도가 빨라졌음을 확인할 수 있었다.

[1]경북 경산시 유곡동 290, 대구한의대학교 정보과학부 교수