준-특이치 분해와 대규모 문서 군집화에의 응용
(A Semi-Singular Value Decomposition and Its Application to Large Text Documents Clustering)

경산대학교 정보과학부
신 양 규

초 록

고차원 희소행렬의 신속하고 정확한 계산은 다양한 분야에 활용된다. 고차원 희소행렬에 특이치 분해를 적용하면 원래 행렬의 구조적 성질을 유지하면서도 불필요한 자료를 제거할 수 있다. 본 논문에서는 대규모 문서 집합의 군집화에 활용할 수 있도록 특이치 분해에서 rank를 필요한 크기만큼 제한하는 준-특이치 분해기법과 이를 이용한 계층적 군집화 기법을 제안한다. 문서 집합을 고차원 희소행렬로 변환하여 준-특이치 분해를 적용하면 군집화를 실시할 때 불필요한 노이즈를 제거시킴과 동시에 초기 클러스터를 생성한다. 초기 클러스터는 neaest neighbor 방법으로 학습을 시킨 후 클러스터 내에 또 다른 클러스터를 연속적으로 생성하는 계층적 군집화 방법이 된다. 이 방법의 장점은 초기 오차를 최소화하면서 원하는 수만큼의 클러스터를 계층적으로 생성할 수 있다는 것이다.