2006, Vol.35, No.5:657-661 ISSN

实际问题研讨

基于EM算法的文本聚类优化研究

冯中慧 鲍军鹏 沈钧毅

(西安交通大学电子与信息工程学院)

Abstract 针对现有的文本聚类算法难以取得满意结果的问题，以EM算法为基础，提出能分别描述相似、不相似聚类对的相似性分布以及重要、不重要文档的重要性分布的文本聚类优化模型（text clustering optimization model, TCOM）. 基于该模型，设计一种通过合并不同的文本聚类结果以获取最优性能的方法. 实验结果表明，利用该方法同时改善了聚类精度和召回率，其性能优于单独使用现有的硬、软聚类算法.

Keywords 硬聚类；软聚类；EM算法；文本聚类优化模型（TCOM）

收稿日期 修回日期

通讯作者

DOI PACS:TP18