An improved ant algorithm with LDA-based representation for text document clustering

1 Introduction

（最后两段）

本文提出了一种用于文档聚类的改进的蚁群聚类算法。在本文提出的聚类方法中，我们利用一种基于蚁群的混合聚类算法 (AntClass) 作为基本模型 [16]。该算法结合了蚁群优化的随机性和探索性特征与 K-means 算法的决定性和启发式特征 [16]。该算法面临的一个问题是，算法终止时所生成的聚类数量与原始数据集中的聚类数量并不相近。该算法趋向于生成更多的聚类。为了增强聚类质量，我们提出了两种用于合并聚类的启发式方法，与 AntClass 算法结合使用。我们将改进后的蚁群聚类算法用于文档聚类，并使用 LDA 进行文档表示。

2 Literature review

2.1 Metaheuristics in clustering

2.2 Latent Dirichlet allocation in clustering

概率主题模型可以被用于概括大型文档集合。对概率主题模型的基本运用是识别文本文档的主题。除了基本运用以外，LDA 和其他概率主题模型方法已被广泛应用于大量的 NLP 应用，包括.....

[41] 检测了概率隐语义分析、概率语义分析以及 K-means 在一系列文本文档中识别主题和主题趋势的性能。实验结果表明，概率隐语义分析是一种可用于在大型语料库中识别有意义的主题的工具。

[42] 提出了一种基于 LDA 的文本表示方法。该方法使用 LDA 来抽取主题。然后，利用同义词词典 (thesaurus) 和基于语料库的相似性度量来计算单词之间的相似度。基于这些相似性，可以获得主题的描述性特征。此外，人为注释提供了基于一致性的主题注释。从主题中获取的主题一致性被用作类别标签。基于单词相似度特征和主题注释，准备了两个数据集。该论文使用分类器 (SVM, K-NN, Random Forest) 对所提出的方法在主题分类上进行了评估。

LDA 方法也能被用于文本文档聚类。本节接下来的内容对此进行了概述。

[43] 提出了一种基于主题模型的文档聚类模型。在该模型中，主题模型被以一种高效的方法用来表示文档，以此降低了文本文档的维度，同时得到了文本的语义关联。

[44] 提出了一种文本文档聚类的三阶段方案。在该方案中，文本文档被表示为 LDA 所取得的主题，并基于主题的意义度 (significance degrees) 识别最有意义的主题。然后，使用 K-means++ 算法确定初始聚类中心。最后，使用 K-means 算法基于潜藏主题实现文档聚类。

[45] 测试了 LDA 及其变式 (the hierarchical LDA, correlated topic models and hierarchical Dirichlet process) 在科学文档聚类上的性能。

[46] 提出了一种基于层次 LDA 的文本文档聚类方法。传统的 bag-of-words 文本文档表示无法取得单词之间的关联信息和共现 (co-occurrence) 关系。因此，[43] 提出的模型利用 LDA 进行文本文档表示。

[47] 提出了一种基于 LDA 和 K-means 的单词聚类方法。在该模型中，LDA 被用于从文本中抽取主题，K-means 的质心则是从概率最高的名词中选取。

[48] 测试了 VSM 和 LDA 在跨语种文档聚类上的性能。

[49] 提出了一种基于 LDA 的多语种主题模型，将现有的方法应用到多种现实任务中。

[50](略)

3 Motivation and contribution of the study

文档聚类的一个重大挑战就是数据的高维问题，为解决难题，现已提出了很多方法，比如获得文本数据集的低维表示，在转换后的空间中工作。例如，非负矩阵因式分解方法可以被用于估计 term-document 矩阵 [51]。无监督的降维方法可以与聚类方法结合使用。例如，4 种降维技术 (independent component analysis, latent semantic indexing, document frequency, random projection) 已经在文档聚类上进行了测试 [52]。[53] 结合了线性判别分析和 K-means 算法，用于文本聚类中的自适应降维。[54] 提出了基于 latent semantic indexing 的方法，用于以语义文本文档的语义一致性表示。[42] 利用 LDA 将文本文档表示为主题的集合。关于元启发式聚类算法在文本领域的最新研究表明，元启发式算法能够在文档聚类上得到理想的结果 [17-27]。虽然 LDA 和元启发式聚类算法的应用备受研究者关注，但是在文本文档聚类中，利用 LDA 来表示文本文档的研究工作非常有限。为了弥补这一空缺，本文采用基于 LDA 的文本表示方案，在文本文档聚类中 25 个广泛使用的文本数据集上进行了实证分析和基准结果分析。此外，本文提出了一种增强的蚁群聚类方案。据我们所知，这是在文本文档聚类中首次对基于 LDA 的文本表示和元启发式聚类方法的全面综合的分析。