doi:10.1631/FITEE.1601125

题目：基于内容和引用的科研文献的主题发现和演化

概要：科研文献数据库中的重要主题随时间的演化的方式已经越来越受到全球研究者的关注。在一个科研论文数据集中，任何一篇论文可以被认为是由组成论文本身的词和论文引用的文献所组成的。在本文中，我们提出了一种名为“Citation-content-LDA (latent Dirichlet allocation)”的主题发现方法，该方法在一个概率生成模型中同时生成文献的引用关系和文献本身的词。Citation-content-LDA 模型利用了一种两层结构的主题模型，即利用引用信息生成父主题和利用文本信息生成子主题。模型参数通过吉布斯采样算法来估计。我们还提出了一个主题演化算法，该算法包括主题分割和主题间依赖关系计算两个步骤。我们在 *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI)和 *IEEE Computer Society* (CS)两个数据集上测试了提出的 Citation-content-LDA 模型和主题演化算法，证明了我们提出的算法能有效的发现重要的主题和反映重要研究主题的主题演化情况。经过我们的评价指标的评测，Citation-content-LDA 算法的性能优于 Content-LDA 和 Citation-LDA 算法。

关键词：主题提取；主题演化；评价方法