基于自适应在线双词主题模型的应用程序评论新兴主题识别

周芃[1]，王勇[1,2]，高翠芸[3]，杨非[4]
[1]安徽工程大学计算机与信息学院，中国芜湖市，241000
[2]南京大学计算机软件新技术国家重点实验室，中国南京市，210000
[3]哈尔滨工业大学（深圳）计算机科学与技术学院，中国深圳市，518000
[4]之江实验室，中国杭州市，310000

摘要：应用程序评论中的新兴主题突出了用户在一定时期内关注的主题（如软件漏洞）。准确、及时地识别新兴主题能帮助开发者更有效地更新应用程序。已有文献基于主题模型或聚类方法识别应用程序评论中的新兴主题。然而，由于评论文本长度较短，提供的信息有限，新兴主题识别准确率较低。为解决该问题，提出一种改进的新兴主题识别方法（IETI）。首先采用自然语言处理技术减少评论文本中的噪音数据，然后使用自适应在线双词主题模型识别评论中的新兴主题。最后利用新兴主题中相关的短语和句子解释新兴主题的含义。采用官方更新日志作为新兴主题的评估标准，选择 6 个常见的应用程序对 IETI 进行评估。实验结果表明，IETI 在识别新兴主题方面优于传统方法，短语标签 F1 值增量为 0.126，句子标签 F1 值增量为 0.061。我们在 Github（https://github.com/wanizhou/IETI）上发布了 IETI 的代码。

关键词：应用程序评论；新兴主题识别；主题模型；自然语言处理
https://doi.org/10.1631/FITEE.2100465