一种新的融合本体和主机信息的改进禁忌搜索算法的主题爬虫方法

刘景发[1]，王震[1,2]，钟国[1]，杨志和[1]
[1]广东外语外贸大学信息科学与技术学院，中国广州市，510006
[2]中国联通中南研究院，中国长沙市，410000

摘要：为解决传统主题爬虫方法存在的主题描述不完整和重复爬取已访问链接的问题，本文提出一种新的融合本体和主机信息的改进禁忌搜索算法的主题爬虫方法（FCITS_OH）。该方法基于形式概念分析（FCA）构建领域本体，在语义和知识层面描述主题。为避免重复爬取已访问的链接和扩大搜索范围，提出一种改进的禁忌搜索（ITS）算法和记忆主机信息的策略。此外，为改进未访问链接的主题相关性的评估方法，提出一种基于Web文本和链接结构的综合优先度评估方法。以旅游和暴雨灾害为主题的实验结果表明，对于不同的性能指标，所提出的爬虫方法优于文献中其它主题爬虫策略。

关键词：主题爬虫；禁忌搜索算法；本体；主机信息；优先度评估
https://doi.org/10.1631/FITEE.2200315