基于本体和模拟退火算法的暴雨灾害主题爬虫策略

刘景发[1,2]，李帆[3]，丁若尧[1,2]，刘子昂[4]
[1]广东外语外贸大学广州市非通用语种智能处理重点实验室，中国广州市，510006
[2]广东外语外贸大学信息科学与技术学院，中国广州市，510006
[3]南京信息工程大学计算机与软件学院，中国南京市，210044
[4]阿尔伯塔大学理学院，加拿大埃德蒙顿市，T6G2H6

摘要：目前，主题爬虫是从海量异构网络中获取有效领域知识的重要方法。目前大多数主题爬虫技术难以获得高质量爬行结果。主要难点包括主题基准模型的建立、超链接主题相关性的评估和爬行策略的设计。本文采用领域本体为特定主题构建主题基准模型，并提出一种新的基于局部本体和全局本体的多重筛选策略（MFSLG）。为提高待访问超链接主题相关性计算精度，提出一种基于网页文本和链接结构的综合优先度评估方法（CPEM），同时，采用模拟退火（SA）算法避免主题爬虫陷入局部最优搜索。本文首次设计融合SA算法、MFSLG策略和CPEM策略实现主题爬虫，提出两种新的基于本体和SA主题爬虫策略（FCOSA），包括基于全局本体的FCOSA策略（FCOSA_G）和基于局部本体和全局本体的FCOSA策略（FCOSA_LG），以从网络中获取与暴雨灾害主题相关的网页。实验结果表明，针对不同性能指标，所提爬虫策略优于其他主题爬虫策略。

关键词：主题爬虫；本体；优先度评估；模拟退火；暴雨灾害
https://doi.org/10.1631/FITEE.2100360