doi:10.1631/FITEE.1601344

题目：基于机器学习的抄袭源检索的查询生成方法

概要：抄袭源检索是抄袭检测的核心任务。使用从可疑文档提取的查询来检索抄袭源已成为抄袭源检索的标准方法。从可疑文档生成查询是源检索最重要的步骤。当前研究主要使用了基于启发式的查询生成方法。然而，每个启发式方法都有其优点，不同方法生成的查询可以获得不同的源检索结果，没有一种方法生成的查询的源检索性能可以在所有的文本片段上具有统计有效性地优于其他方法。这使得基于启发式的源检索查询生成方法的性能改善主要依赖专家经验。因此，很难开发一种可以克服现有启发式方法的新方法。本文提出使用统计机器学习方法解决源检索的查询生成问题，将源检索的查询生成形式化到一个排序学习的框架下，从备选查询中选择有利于提高源检索性能的查询，力争在每个可疑文档片段上获得最优的源检索性能。据我们所知，这是第一项应用机器学习方法解决源检索查询生成问题的工作。为了解决排序学习训练用例的缺失，提出了基于现有源检索语料构建查询生成语料的方法。在 PAN 抄袭源检索评测数据上的试验结果证明了该方法具有统计意义地优于多个基线方法。

关键词：抄袭检测；源检索；查询生成；机器学习；排序学习