摘 要

网络信息的迅猛发展对网络信息检索技术的进步提出了空前迫切的要求。网络信息检索所采用的“关键词查询+选择性浏览”的信息交互方式便利了用户获取信息的过程，但也带来了用户查询信息需求传递的瓶颈问题，造成了检索技术发展的障碍。本文尝试对检索用户群体行为进行多角度和多层次的分析，以便更准确地理解用户需求，从而提高网络信息检索系统的性能。本文的贡献主要集中在以下几个方面：

第一，从网络数据处理的角度，提出一种基于用户群体需求分析的网络数据质量评估方法。该方法将能否满足用户查询需求作为网页质量评判的依据，通过分析网页集合的特征，利用贝叶斯方法计算网页成为高质量页面的可能性。在海量网络语料（其规模超过中文互联网网页总数的 4%）上进行的网页质量评估实验和在 TREC 网络信息检索评测平台上进行的检索实验结果证明，此方法能够在减少 95%页面索引规模的情况下保留 90%以上的高质量页面，实现在减少检索系统索引量的同时提升检索性能。

第二，从用户查询处理的角度，提出一种基于用户行为分析的查询信息需求分类方法。该方法基于对用户查询点击群体行为的分析方法识别个体用户查询需求类别。基于大规模真实搜索引擎用户日志的实验说明，此方法能够对超过 80%的用户信息需求正确分类，算法性能比传统方法有 20%左右的提高。

第三，针对网络信息检索系统的评价问题，基于用户海量搜索日志挖掘提出一种新的搜索引擎评价的自动化处理方法。该方法在通用的信息检索系统评价框架下，用客观的面向用户日志挖掘的自动化方法代替主观的人工评价。实验结果证明，此方法能与人工标注的评价取得基本一致的评价效果，同时大大减少了评价所需的人力、物力资源，并加快了评价反馈周期。

第四，综合用户群体行为分析，提出了一种网络信息检索系统架构的改进方案。该方案设计面向网络数据对象处理的层次索引结构，以基于宏观用户行为分析的查询处理模型为核心，综合本文研究成果，实现在当前的检索系统交互方式限制下尽量缓解用户信息需求传递的瓶颈问题。

关键词：网络信息检索；用户群体行为分析；网页质量评估；检索性能评价